Australian Bureau of Statistics

1351.0.55.054 - Research Paper: Big Data, Statistical Inference and Official Statistics, Mar 2015

Latest ISSUE Released at 11:30 AM (CANBERRA TIME) 18/03/2015 First Issue

Summary

Executive Summary

EXECUTIVE SUMMARY

Official statisticians have been using a diversity of data sources in the production of official statistics for decades, including “designed” data sources such as censuses and surveys, and “found” data sources such as administrative and transactional data.

As a result of more and more interaction with digital technologies by citizens, and the increasing capability of these technologies to provide digital trails, new sources of data have emerged and are increasingly available to official statisticians. Such sources include data from sensor networks and tracking devices e.g. satellites and mobiles phones, behaviour metrics e.g. search engine queries, and on-line opinion e.g. social media commentaries. The collective term for such data sources is Big Data.

Whilst Big Data have the potential to create a rich, dynamic and focussed picture of Australia for informed decision making, and to improve the efficiency in the production of official statistics, this paper contends that there are a number of issues that an official statistician has to consider before deciding if a particular source from Big Data can be used for the regular production of official statistics.

A principal decision is business need and business benefit. This includes consideration of whether the new data source will improve the offerings of an existing statistical series, or plug statistical data gaps e.g. increasing the frequency of release, improving the richness of details such as small area or small population group statistics, or providing new official statistics that cannot be cost effectively provided using existing data sources. It also includes assessment of the business case in using the new data source, such as whether there will be a reduction of cost in the statistical production or reduction in provider load, and assessment of the quality of statistics produced from Big Data using Data Quality frameworks, against the benefits to be provided from the new source.

Another key decision is the validity of statistical inferences from Big Data. Big Data, depending on the source, suffer from one or more statistical biases, e.g. coverage bias, representational bias or self-selection biases, and measurement errors. Unlike errors due to sampling, the magnitude of these types of error will not be reduced by increasing the size of the data set.

The challenge for official statisticians is to develop a suitable methodology for analysing such data sets so that any conclusions drawn from the analysis are valid statistically. Firstly, official statisticians need a methodology to address any bias from Big Data, and secondly, a methodology in using Big Data to produce fit-for-purpose official statistics.

A Bayesian inference framework is adopted in this paper to assess the conditions under which valid statistical inference can be drawn from Big Data. The conditions are similar to those for making valid statistical inference from survey data: that any underlying process for the inclusion or exclusion of information from the Big Data source is independent of that information *per se*.

By treating Big Data as auxiliary information, and integrating census and survey data – ground truth data – with Big Data, this paper also provides a Bayesian method for using new data sources to produce official statistics. For count data, a dynamic logistic regression model is used. For continuous data, a dynamic linear model is described. The dynamic logistic model is applied to the theoretical analysis of satellite imagery data for the prediction of crop growing areas in Australia.

Other relevant issues for the official statistician to consider when deciding if a particular source from Big Data is to be used for the production of official statistics are: privacy and public trust, data ownership and access, computation efficiency and technology infrastructure.

Until recently, the Australian Bureau of Statistics' (ABS) progress in Big Data domain has been primarily review and monitoring of industry developments while contributing to external strategic and concept development activities. This paper summarises the ABS Big Data Strategy with objectives to build an integrated multifaceted capability for systematically exploiting the potential value of Big Data for official statistics.

This paper also describes the ABS Big Data Flagship Project, which has been established to provide the opportunity for the ABS to gain practical experience in assessing the business, statistical, technical, computational and other issues related to Big Data as outlined earlier in this paper. In addition, ABS participation in national and international activities on Big Data will help it share experience and knowledge, and collaboration with academics will help ABS better acquire the capability addressing business problems using Big Data as a part of the statistical solution.

About this Release

Whilst Big Data have the potential to improve the statistical production and statistical offerings, this paper outlines the issues that need to be considered by the official statistician, before a particular Big Data source can be used for the regular production of official statistics. In addition, the paper outlines Bayesian methods for analysing Satellite imagery data, and also the ABS strategies and initiatives on Big Data.